BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26369
doi: 10.3897/biss.2.26369

OPEN ACCESS

Conference Abstract

Best practices for connecting genetic records with specimen data

Michael Trizna ‡

‡ Smithsonian Institution, Washington, DC, United States of America

Corresponding author: Michael Trizna (mike.trizna@gmail.com)

Received: 02 May 2018 | Published: 03 Jul 2018

Citation: Trizna M (2018) Best practices for connecting genetic records with specimen data. Biodiversity Information Science and Standards 2: e26369. https://doi.org/10.3897/biss.2.26369

Abstract

As rapid advances in sequencing technology result in more branches of the tree of life being illuminated, there has actually been a decrease in the percentage of sequence records that are backed by voucher specimens Trizna 2018b. The good news is that there are tools Trizna (2017), NCBI (2005), Biocode LLC (2014) to enable well-databased museum vouchers to automatically validate and format specimen and collection metadata for high quality sequence records. Another problem is that there are millions of existing sequence records that are known to contain either incorrect or incomplete specimen data. I will show an end-to-end example of sequencing specimens from a museum, depositing their sequence records in NCBI's (National Center for Biotechnology Information) GenBank database, and then providing updates to GenBank as the museum database revises identifications. I will also talk about linking records from specimen databases as well. Over one million records in the Global Biodiversity Information Facility (GBIF) Trizna (2018a) contain a value in the Darwin Core term "associatedSequences", and I will examine what is currently contained in these entries, and how best to format them to ensure that a tight connection is made to sequence records.

This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

Keywords

collections, GenBank, museum vouchers, biodiversity informatics, software, data linking, Darwin Core

Presenting author

Michael Trizna

References

- Biocode LLC (2014) Biocode Field Information Management System. https://github.com/biocodellc/biocode-fims
- NCBI (2005) Using NCBI LinkOut. https://www.ncbi.nlm.nih.gov/books/NBK3810/
- Trizna M (2017) Genetic Collections Python library. https://github.com/MikeTrizna/genetic_collections
- Trizna M (2018a) iDigBio associatedSequences count. https://gist.github.com/MikeTrizna/f98cc3e4c7c393f043fecadeccf465c1
- Trizna M (2018b) Voucher counts in GenBank. https://gist.github.com/MikeTrizna/c1fe18ceda7531fedba730acc528d588